社会计算

第六节

经典传播模型

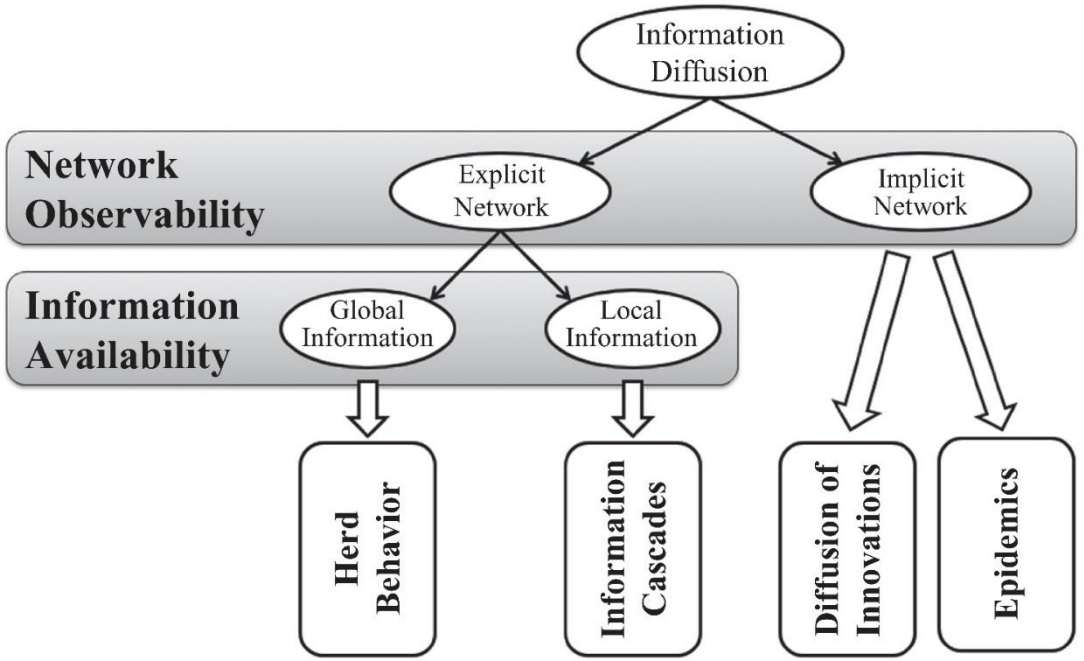
徐童

2024.4.8

上节回顾

- **如何描述传播过程**

- 前一节中信息级联的例子，为我们描述了基本的信息扩散过程
- 如何为这种信息扩散的过程提供基于网络结构的量化建模?
- 本节课中，我们将介绍最后一类：信息级联模型。需要注意：
  - 局部信息可达的约束在现实中较难实现
  - 经典模型往往对问题进行了较大的简化



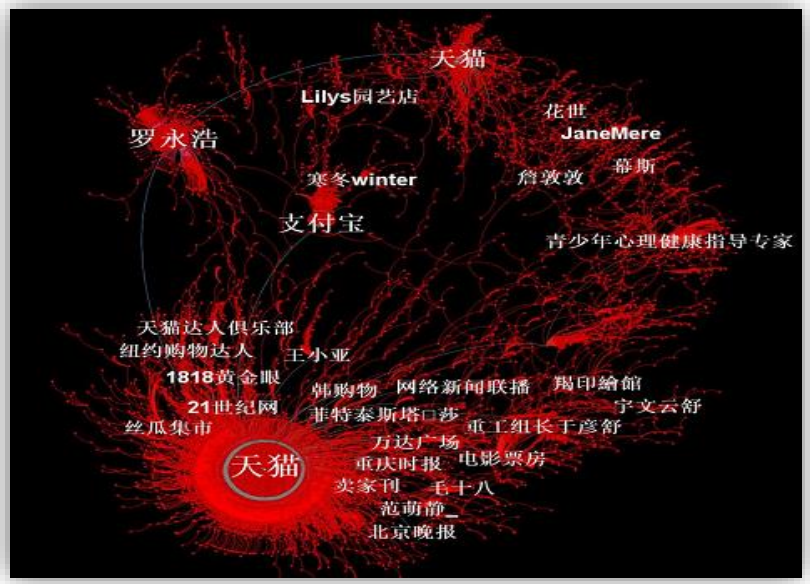
本节目录

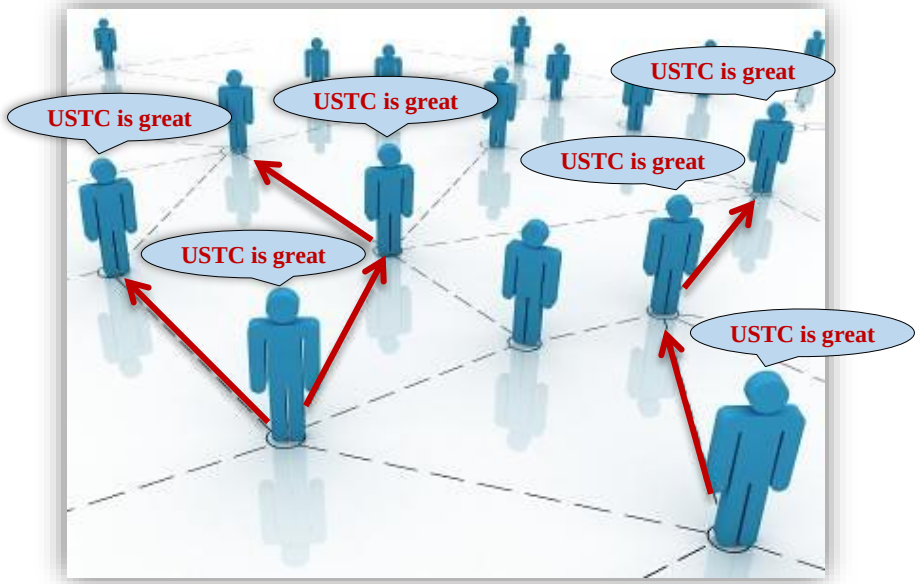
- **基本传播模型**
  - 独立级联模型
  - 线性阈值模型
- 传播最大化问题
- 衍生传播模型与传播问题
- 基于社会网络的推荐

本节主旨

- **依托信息传播的口碑营销**

- 大量实践已经证实，相较于普惠型营销，口碑营销成本更低，传播也往往更快
  - 在这一过程中，“大V”们扮演着至关重要的角色





• 信息传播的元素

- 一般而言，我们将信息传播过程中涉及的元素归为以下三类：
  - 发送者（Sender），也称作信息源（Source）或“种子节点”（Seed）
    - 指在信息传递开始时拥有信息的那一小部分用户集合
  - 接收者（Receiver），指作为潜在传播目标的广大用户集合
    - 接收者集合的规模要远大于前者，且不同发送者的目标集合存在重叠
  - 媒介（Medium），指传播过程发生的平台
    - 例如，寻找红气球比赛中的社交媒体 / 论坛、QQ群等

- **信息级联中的基本假设**

- 首先，信息级联发生在一张有向图上
  - 对于无向图，可以将其连边转化为双向边进行处理
  - 对于网络中的节点来说，这些边就是信息传递的媒介
- 其次，每个节点仅能将信息传递给与其直接相连的节点
  - 例如，大V可以将信息传递给其粉丝，但不能传递给未关注他/她的人
    - 信息传递的局部可达性！

- **信息级联中的基本假设**

- 特别需要注意的是，网络中节点的状态是二元（Binary）的
  - 激活（Active/Activated）：表示节点已经收到了这一信息
  - 未激活（Inactive）：表示节点尚未接收到这一信息
    - 不存在薛定谔的状态！

➢ 一个小问题：什么情况下算是激活？

- 接收到信息，并且尝试将信息传给别人才叫激活
  - 两个动作缺一不可（是否合理？）

- **信息级联中的基本假设**

- 特别需要注意的是，网络中节点的状态是二元（Binary）的
- 已激活的节点才具备激活其他节点的能力
  - 而且，激活能力有一定的**时限**！
    - 传播中存在着时间“轮次”的概念
    - 类似于传染病模型SIR的设定
    - 这个设定是否普遍合理？如何确定一个合理的时限？
      - 后面我们会展示去除这一约束的特殊模型

- **信息级联中的基本假设**

- 特别需要注意的是，网络中节点的状态是二元（Binary）的
- 激活是**不可逆**的过程
  - 可以从未激活到激活，但不能从激活退回未激活
  - 这个约束又是否普遍合理？
    - 核心争议在于：是否接受信息的“二次传播”
    - 后面我们同样会展示去除这一约束的特殊模型

- **基本模型（1）独立级联模型**

- 独立级联模型（Independent Cascade Model）
  - “独立”体现在，每次激活都是一次独立事件，相互不产生影响
    - 激活的尝试相当于一次以特定概率抛硬币的过程
  - 同时，每个已激活节点，只有一次机会尝试激活他/她的未激活邻居节点
    - 一旦尝试失败，不会再有第二次尝试机会

- **基本模型（1）独立级联模型**

- 独立级联模型（Independent Cascade Model）中的重要概念：轮次
  - 如果某个节点在第 t 轮被激活，那么，他仅有一次机会，即仅能在 t+1 轮，尝试激活他所有未被激活的邻居节点
    - t = 1时，仅有种子节点可以尝试激活其他节点
  - 对于节点 v 而言，他激活邻居节点 w 的概率采用 P_{vw} 表示
    - 以 P_{vw} 为概率进行抛硬币
  - 整个传播过程直到所有节点都被激活，或没有新节点可以被激活为止

- **基本模型（1）独立级联模型**

- 独立级联模型（Independent Cascade Model）中的重要概念：轮次
  - 对于节点 v 而言，他激活邻居节点 w 的概率采用 P_{vw} 表示
    - 以 P_{vw} 为概率进行抛硬币
    - P_{vw} 的取值方式：
      - 基本传播模型里，为简化考虑，一般将P_{vw}设为1/N，N为w节点的入边的数量
      - 当然，也有实现确定带权图的做法（如后续的例子）
      - 此外，也可以基于主题等因素对P_{vw}进行扩展

- **基本模型（1）独立级联模型**

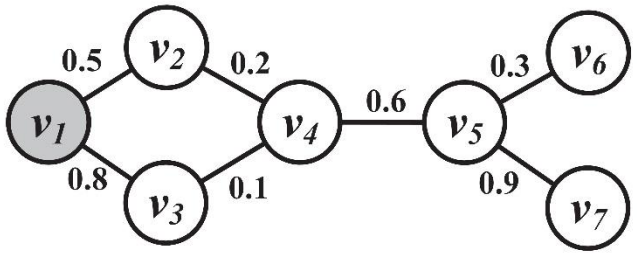
Algorithm 7.1 Independent Cascade Model (ICM)

Require: Diffusion graph $G(V, E)$, set of initial activated nodes A_0, activation probabilities $p_{v,w}$

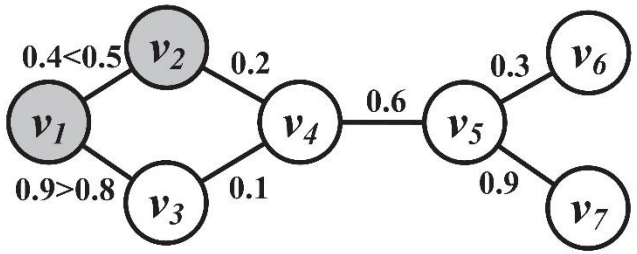
return Final set of activated nodes A_∞
i = 0;
while $A_i \neq \{\}$ **do**

  $i = i + 1$;
  $A_i = \{\}$;
  for all $v \in A_{i-1}$ **do**
    for all w neighbor of v, $w \notin \cup_{j=0}^{i} A_j$ **do**
      rand = generate a random number in [0,1];
      if rand < $p_{v,w}$ **then**
        activate w;
        $A_i = A_i \cup \{w\}$;
      end if
    end for
  end for
end while
$A_\infty = \cup_{j=0}^{i} A_j$;
Return A_∞;

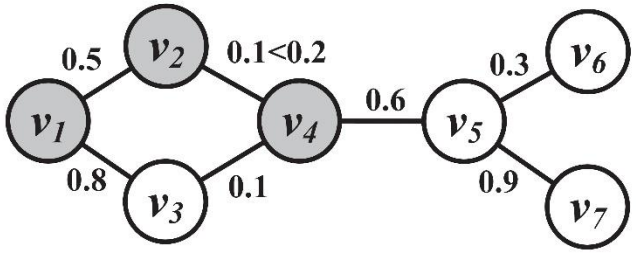
- **基本模型（1）独立级联模型**



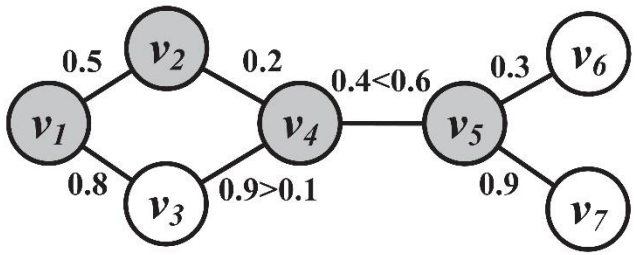
Step 1



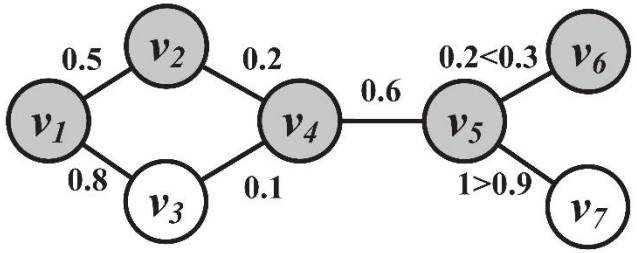
Step 2



Step 3



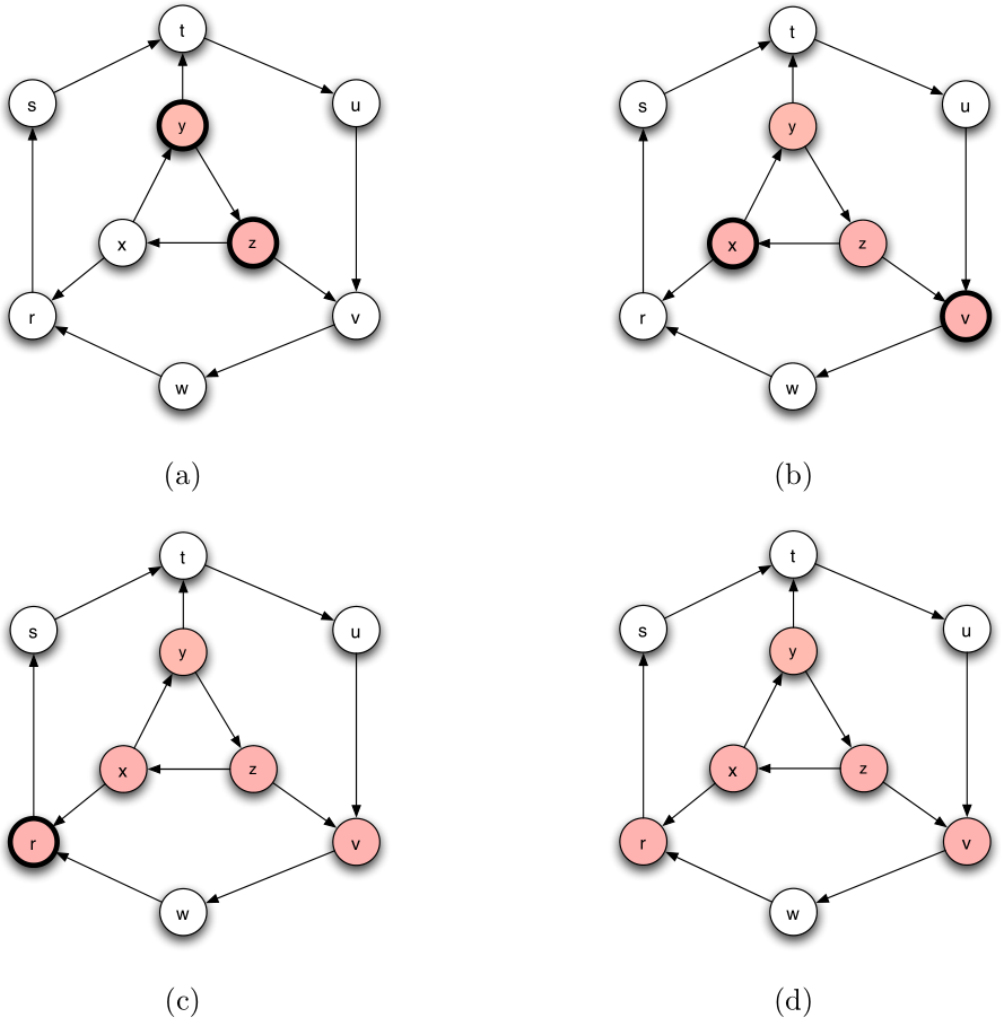
Step 4



Step 5

• 小复习：SIR与独立级联模型的关联

- 在SIR中，有类似的传播过程
  - 如果某条边被选中，则视作“开放边”
  - 反之，则视作“阻塞边”
  - 因此，某个节点被传染，当且仅当它到种子节点有一条全部由开放边组成的路径
    - 可被传染的节点的集合被称作Reverse Reachable Set（RR Set）
    - 在独立级联模型中，我们有类似的做法和类似的定义



- **基本模型（2）线性阈值模型**

- 线性阈值模型（Linear Threshold Model）
  - 另一种视角：将信息传递过程视作多人影响的叠加过程
  - 一个用户会被某个信息激活，如果来自他已激活邻居的影响超过某个阈值

$$\sum_{v_j \in N_{\text{in}}(v_i), v_j \in A_{t-1}} w_{j,i} \geq \theta_i$$

    - 阈值预先设定，往往为从[0,1]均匀分布中随机抽取的一个数值
      - 更复杂的情况：可以根据用户对信息的兴趣等决定

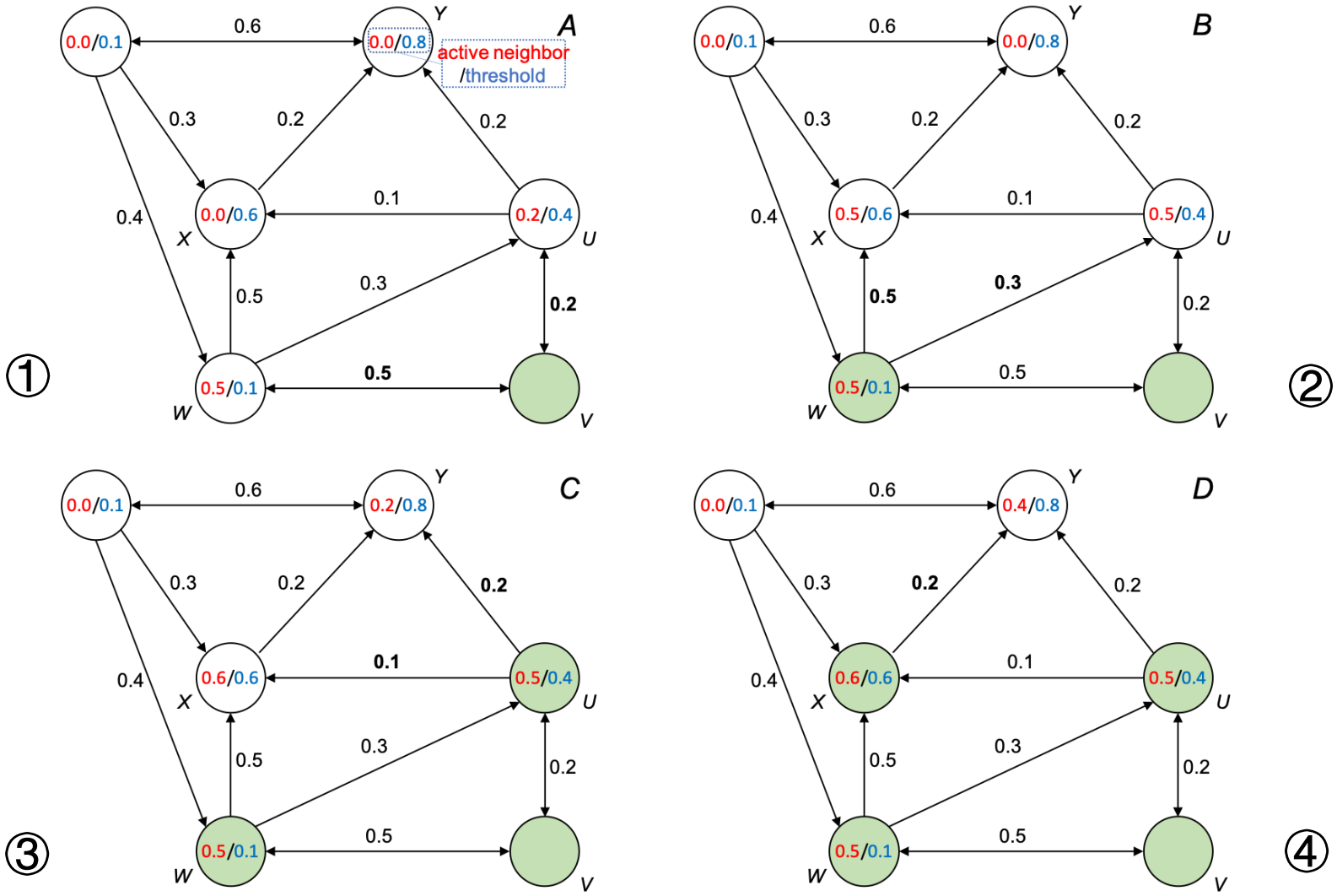
- **基本模型（2）线性阈值模型**

Algorithm 8.1 Linear Threshold Model (LTM)

Require: Graph $G(V, E)$, set of initial activated nodes A_0

1: **return** Final set of activated nodes A_∞
2: i=0;
3: Uniformly assign random thresholds θ_v from the interval [0, 1];
4: **while** $i = 0$ or $(A_{i-1} \neq A_i, i \geq 1)$ **do**
5: $A_{i+1} = A_i$
6: inactive $= V - A_i$;
7: **for all** $v \in$ inactive **do**
8: **if** $\sum_{j \text{ connected to } v, j \in A_i} w_{j,v} \geq \theta_v$. **then**
9: activate v;
10: $A_{i+1} = A_{i+1} \cup \{v\}$;
11: **end if**
12: **end for**
13: $i = i + 1$;
14: **end while**
15: $A_\infty = A_i$;
16: Return A_∞;

• 基本模型（2）线性阈值模型



• 基本模型（2）线性阈值模型

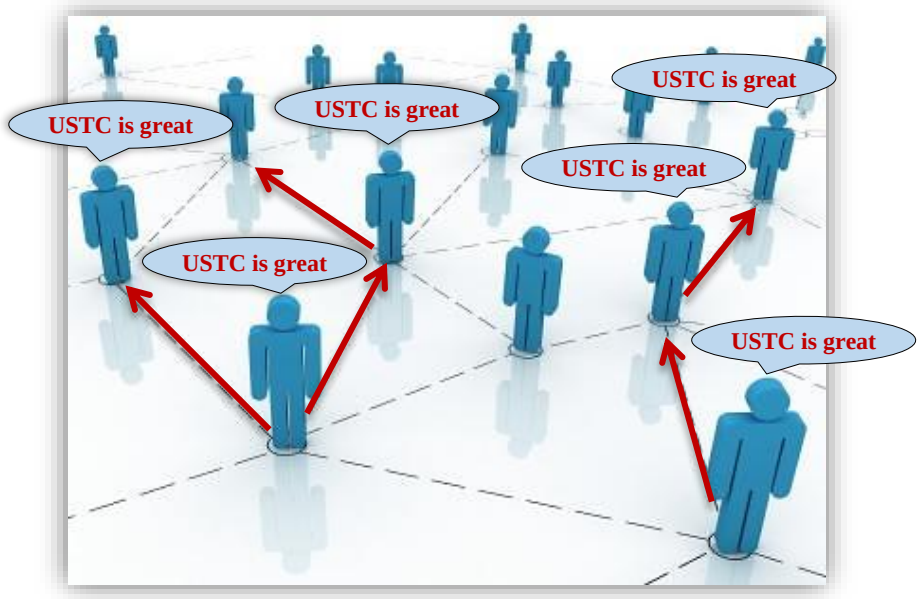
- 线性阈值模型与独立级联模型的区别：随机性
  - 对于独立级联模型来说，其随机性在于抛硬币的过程
    - 因此，独立级联模型是完全随机过程，每一次的结果可能都不相同
    - 一般需要重复多次以确定个体节点被激活的可能性
  - 对于线性阈值模型来说，其随机性在于边权重/阈值的确定
    - 如果采用启发式方法确定边权/阈值，则该方法结果完全由方法设计决定
    - 一旦确定边权/阈值（无论何种方式），其结果具有唯一性

本节目录

- 基本传播模型
  - 独立级联模型
  - 线性阈值模型
- **传播最大化问题**
- 衍生传播模型与传播问题
- 基于社会网络的推荐

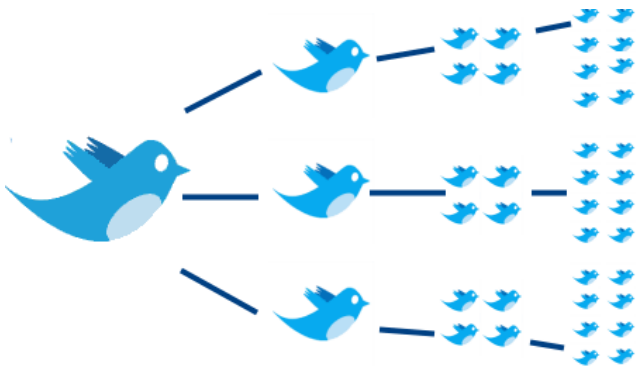
- **社会网络中的信息传播**

- 口碑营销（Word of Mouth）
  - 信息传播的过程，核心在于信息的接受
    - 传统的信息传播建模，往往将信息传播与信息接受合二为一，传播即视作接受了信息
    - 因此，信息传播分析在市场营销领域有着大量的应用



- **信息传播最大化问题**

- 为什么会有传播最大化问题?
  - 口碑营销的常见应用：通过优惠来吸引潜在客户
    - 例如，通过发放优惠券 / 赠品的方式来扩大客户群
  - 然而，商家的预算是有限的
    - 因此，往往仅能通过收买少数用户来扩散消息
    - 这个时候，选择目标用户就至关重要!



- **信息传播最大化的启发式方法**

- 解决传播最大化问题的启发式思路：寻找网络中最具影响力的节点
  - 例如，如果你想宣传自家商品，找网红带货是个不错的手段
  - 找到影响力节点后，由他们发起信息传播
  - 问题在于，谁是最具影响力的节点?

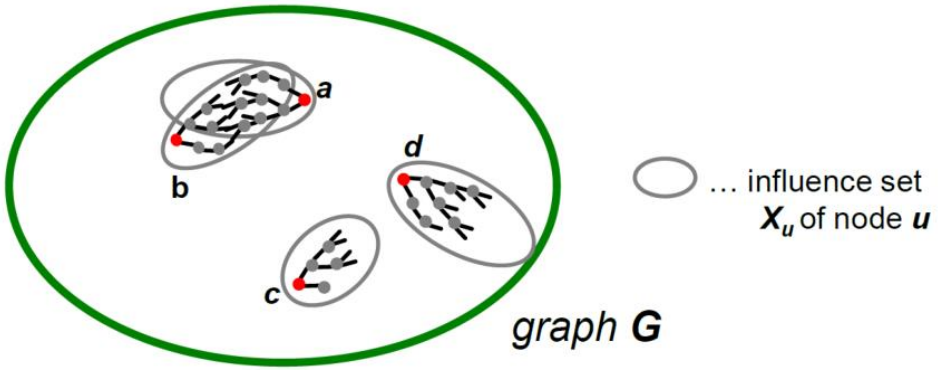
- **信息传播最大化的启发式方法**

- 启发式方法（1）PageRank及其衍生模型
  - 在网页排序部分，我们曾经介绍过，PageRank可用来衡量网页权威性
    - 因此，PageRank及其各种衍生算法如HITS都可以采用

- 启发式方法（2）核心性（Centrality）度量
  - 用于衡量网络中最重要的节点。常见核心性度量如度（Degree）、紧密度（Closeness）、介数（Betweenness）等

- **信息传播最大化的启发式方法**

- 启发式方法（3）计算单个节点所能够激活的邻居数量，再进行排序
- 上述启发式方法，在寻找“最具影响力的节点”时可行
- 然而，在确定影响力节点集合时不可行
  - 节点的影响范围之间可能存在重叠
    - 在单个节点影响力够强的情况下，没必要重叠“双保险”



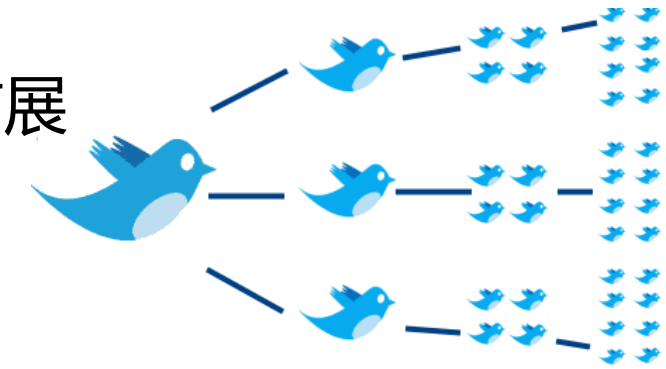
- **信息传播最大化问题**

- 由此，我们正式给出传播最大化问题（Propagation Maximization）的定义
- 假定初始的种子节点集合为 S，预期激活的节点集合为 $f(S)$
- 信息传播最大化的目的，在于限定 S 集合的前提下，最大化 $f(S)$ 的规模
  - 常见的约束为限定 S 集合的规模，即 |S|= k （初始节点集合规模限定为k）

$$\max_{S \text{ of size } k} f(S)$$

  - 如果 S 集合中的节点价值不等，则可将约束进一步扩展

Y Yang, et al., Continuous Influence Maximization: What Discounts Should We Offer to Social Network Users?, SIGMOD 2016



- **一般化的信息传播最大化方法**

- *f(S)*的一些有趣的性质：子模特性（Submodularity）
- 1. *f(S)* 函数是非负的（显而易见）
- 2. *f(S)* 函数是单调非减的，即 $f(S+v) \geq f(S)$
  - 也很好理解，新增加一个节点，至多不增加新激活，不至于减少
- 3. *f(S)* 函数是具有子模特性（Submodularity）的，即：
  - 对于任何集合对S，T，且满足S ⊆ T时，给定节点 v，有

$$f(S \cup \{v\}) - f(S) \quad \geq \quad f(T \cup \{v\}) - f(T),$$

- **信息传播最大化问题的解决方案**

- 基于前述问题定义，我们有坏消息，也有好消息
- 坏消息：在ICM / LTM等模型定义下，传播最大化问题是个NP难问题
  - 简要证明思路：将这两个模型定义下的传播最大化问题，归约为集合覆盖（Set Cover）和节点覆盖（Vertex Cover）问题
    - 详细证明可参见如下论文：

D. Kempe, et al., Maximizing the Spread of Influence through a Social Network, KDD 2003

- **信息传播最大化问题的解决方案**

- 基于前述问题定义，我们有坏消息，也有好消息
- 好消息：由于 $f(S)$ 函数具有子模特性，我们可以采用贪心算法近似求解
  - 以空集合为起点，即初始 S = ∅
  - 经过 k 次迭代，每次选择最大化 $f(S \cup \{v\}) - f(S)$ 的节点 v
  - 效果如何？论文证实贪心算法可以实现至少 $(1 - 1/e)$ 的近似效果
    - 这就意味着，贪心法所得 S 可以激活至少 63% 最优解能激活的节点数

G. Nemhauser, et al. An analysis of the approximations for maximizing submodular set functions. Mathematical Programming, 14(1978), 265–294.

- **信息传播最大化问题的解决方案**

- 基于贪心算法的传播最大化问题求解伪代码：

Algorithm 7.2 Maximizing the spread of cascades – Greedy algorithm

Require: Diffusion graph $G(V, E)$, budget k

1: **return** Seed set S (set of initially activated nodes)
2: $i = 0$;
3: $S = \{\}$;
4: **while** $i \neq k$ **do**
5: $v = \arg\max_{v \in V \setminus S} f(S \cup \{v\})$;
or equivalently $\arg\max_{v \in V \setminus S} f(S \cup \{v\}) - f(s)$
6: $S = S \cup \{v\}$;
7: $i = i + 1$;
8: **end while**
9: Return S;

本节目录

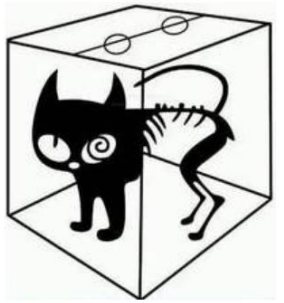
- 基本传播模型
  - 独立级联模型
  - 线性阈值模型
- 传播最大化问题
- **衍生传播模型与传播问题**
- 基于社会网络的推荐

- **独立级联模型的局限性**

- 独立级联模型具有易于求解，假设直观的优点，但也存在一些缺点
  - 小问题 1：每个节点只有一次传播信息的机会，是否过于苛刻？
    - 实际情况下，只要信息还在，就可以持续输出影响
  - 小问题 2：节点状态未必二元化，也难以获得清晰明确的激活轮次
    - ICM 适合类似微博等具有明确转发记录的场景
    - 然而，很多场景下并没有明确的信息传播轨迹
      - 节点是否真的被激活？何时被激活？无法回答

- **稳定状态传播模型**

- 独立级联模型的松弛版本：稳定状态传播（Steady State Spread，SSS）
  - 对ICM的改动体现在以下两点
    - 节点状态不再二元化，而是引入一个变量表示当前被激活的概率
      - 薛定谔的节点出现了！
    - 如果被激活概率不为0，则节点可以持续对外输出信息 / 影响



C Aggarwal, et al., On Flow Authority Discovery in Social Networks, SDM 2011

- **稳定状态传播模型**

- 独立级联模型的松弛版本：稳定状态传播（Steady State Spread，SSS）
  - 稳定状态传播模型的核心公式：

$$I_t(u) = 1 - \Pi_{v \in N(u) \cap I_{t-1}(S)} (1 - p_{v,u}).$$

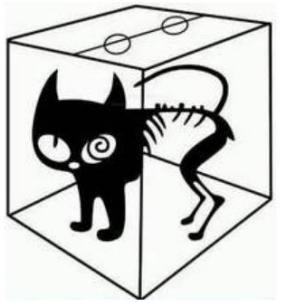
基本的独立级联模型公式 个

$$1 - \pi(i) = \prod_{l \in N(i)} (1 - \pi(l) \cdot p_{li})$$

注意此处

$\Pi(l)$ 表示 l 节点的当前状态，即被激活的概率

当 $\Pi(l) = 0$时，显然不影响邻居节点激活状态

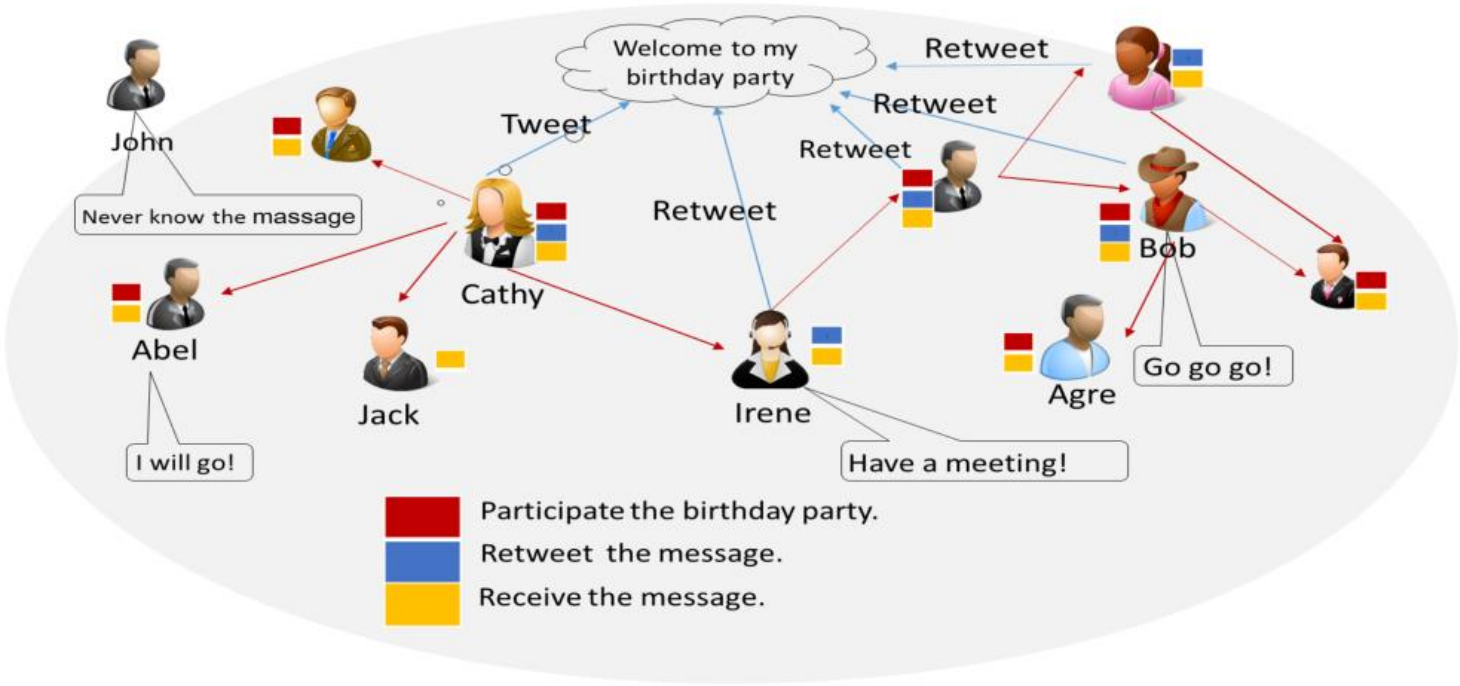


- **信息级联的另一个小问题**

- 独立级联模型在场景应用上还有另一个局限性：信息传播与接收的捆绑
  - 回想一下，ICM / LTM等模型在假设有一个前提，一个转发行为 = 传播 + 接受，两者缺一不可
    - 但实际上，两者不可混为一谈
    - 例如，有些人可能看到并接受了信息，但由于种种原因并没有转发；而另一些转发的人可能实际上并没有接受信息

- **信息级联的另一个小问题**

- 信息传播与接收不一定捆绑的一个实例



- **信息级联的衍生模型**

- 实现信息传播与接收的解绑，先从目标函数改起
  - 第一种思路：考虑信息覆盖问题，即信息覆盖了多大的人群
  - 核心假设：如果某个节点被激活，那么他的所有邻居都被覆盖
  - 由此，衍生出信息覆盖最大化问题

$$\arg\max_{S} \; F(S) = E(|I(S)|) + E(|\bigcup_{a \in I(S)} N(a)|)$$

$$s.t. \quad |S| = k$$

邻居部分

- **信息级联的衍生模型**

- 信息覆盖最大化问题同样具有信息传播最大化问题的性质
  - 因此，可以采用类似的贪心算法加以求解
  - 详细证明与算法可参见如下论文：

    Z Wang, et al., Maximizing the Coverage of Information Propagation in Social Networks, IJCAI 2015

  - 然而，这篇论文仍有进一步拓展的空间：收到信息 =？ 接受信息
    - 实际上，大多数我们收到的信息都被忽略了

- **信息级联的衍生模型**

- 更进一步实现信息传播与接收的解绑，修改模型框架
  - 第二种思路：单独对信息接受过程进行建模

$$F(S) = \mathbf{Adopt}(S) = \sum_{u \in V} [f_u(A_u)]$$

  - 其中，引入函数 $f_u(A_u)$，描述 u 节点接受信息的概率
    - 显然，这一概率与有多少个邻居已经接受了信息正相关
    - A_u 的定义：$u \cup N^{in}(u) \cap \mathbf{Active}(S)$

- **信息级联的衍生模型**

- 更进一步实现信息传播与接收的解绑，修改模型框架
  - 第二种思路：单独对信息接受过程进行建模
  - 由此，衍生出第二个新问题：信息接受最大化问题
    - 即，什么种子节点集合会导致接受信息的节点数期望最大
    - 该问题在$f_u(A_u)$符合一定特性时，同样与信息传播最大化具有类似属性
    - 相关详细证明与算法可参见如下论文：

T Jin, et al., Maximizing the Effect of Information Adoption: A General Framework, SDM 2018

- **信息级联的衍生模型**

- 事实上，信息接受最大化问题可视作一个更为一般化的框架
  - 当满足如下条件时，该问题可退化为信息传播最大化问题：

$$f_v(A_v) = \begin{cases} 1 & \text{if } v \in A_v \\ 0 & \text{if } v \notin A_v. \end{cases}$$

  - 当满足如下条件时，该问题可退化为信息覆盖最大化问题：

$$f_v(A_v) = \begin{cases} 1 & \text{if } A_v \neq \emptyset \\ 0 & \text{if } A_v = \emptyset. \end{cases}$$

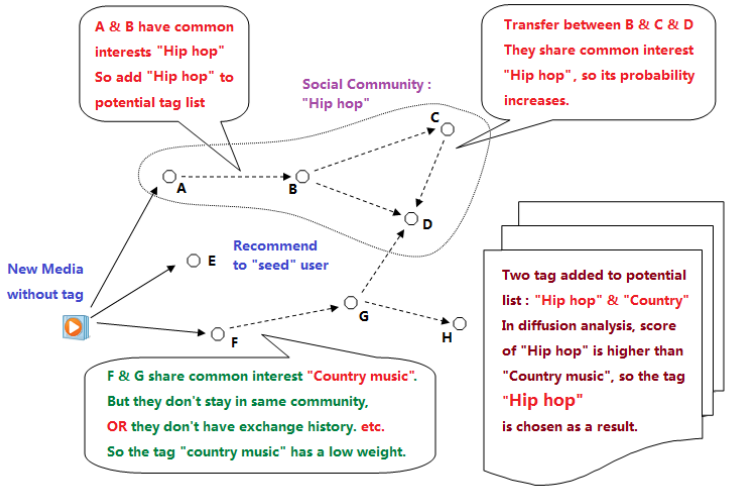
- **主题敏感的传播/互动行为**

- 在线社区尤其是主题相关的社区中，用户的传播行为也体现出鲜明的主题性
  - 共同兴趣在用户社交互动行为中承担了“中介”的作用。

  - **主题敏感性**
    - 共同兴趣与信息主题相契合的两两用户之间更可能发生互动。
  - **局部稠密性**
    - 具有共同兴趣的社团内部的互动行为相比于全局更为稠密。

- **主题敏感的传播/互动行为**

- 分析主题敏感的传播行为，能够为我们提供哪些可能?
  - 主题敏感传播，有助于揭示用户之间的共同兴趣
    - 由此，可以借助社交推荐实现更有效的信息推荐和覆盖
  - 主题敏感传播，还可以用于揭示被传播内容的主题
    - 偶然的点击或无意义，但频繁传播与互动一定喜欢



- **主题敏感的传播/互动行为**

- 如何利用和量化描述这种主题敏感传播?
  - 针对某一社交媒体信息所产生的传播图结构 ***G***，是由用户共同兴趣 ***C*** 与媒体信息主题 ***T*** 共同作用下的结果。
  - 准确还原用户共同兴趣与媒体信息主题两个要素，可实现历史传播图的**“重现”**。
    - 注意：用户的共同兴趣直接落在边上，而不是节点的属性

$$(T_i^*, C^*) = \arg\max_{(T_i, C)} P(G_i | T_i, C)$$

- **主题敏感的传播/互动行为**

- 如何利用和量化描述这种主题敏感传播？
  - 相应的，针对上述两个要素的各自优化，将可得到两个不同的优化目标，对应有监督学习中的**训练**与**测试**阶段
    - 训练阶段：基于已知主题的传播，推测用户间的共同兴趣

$$C^* = \arg\max_{C} \sum_{m_i \in M} D(G_i, T_i, C|_{E_i}).$$

    - 测试阶段：基于已知的用户共同兴趣，推测新传播的内容主题

$$T^* = \arg\max_{T} D(G_i, T_i, C|_{E_i})$$

- **主题敏感的传播/互动行为**

- 如何求解我们先前提出的两个优化问题？
  - 引入信息级联模型，描述传播过程（主要挑战在于主题信息如何引入）
  - ➢ 基于共同兴趣的传播最大化问题
    - 在原传播图中进行主题敏感传播，如能最大化覆盖传播图中的所有节点，即视作“重现”原传播过程
    - 将用户之间的主题敏感互动行为表征为社交传播行为：

$$P_r(k) = 1 - \prod_{u_s \in N(r)} \left(1 - w_{sr} P_s(k-1)\right),$$

- **主题敏感的传播/互动行为**

- 如何求解我们先前提出的两个优化问题？
  - 引入信息级联模型，描述传播过程（主要挑战在于主题信息如何引入）
  - ➢ 基于共同兴趣的传播最大化问题
    - 更进一步，需要考虑w的取值问题
    - 用户之间发生主题敏感互动行为的概率，即社交传播中的“激活概率”，由共同兴趣与信息主题共同决定：

$$w_{sr}^{i} = 1 - \prod_{z=1}^{|T|} (1 - c_{sr}^{z} \cdot a_{i}^{z}).$$

- **主题敏感的传播/互动行为**

- 最终，基于上述量化建模来描述主题敏感传播过程
  - 训练阶段：

$$\min \sum_{s,r,i:e_{sr}\in E_i, m_i \in M} \left(\prod_{z=1}^{|T|}(1 - c_{sr}^{z} \cdot a_{z}^{i})\right),$$

$$s.t. \sum_{z=1}^{|T|} c_{sr}^{z} = 1.$$

  - 测试阶段：基于贪心算法的传播最大化实现
  - 实验证实该方法相较对比算法有20-50%的提升，并在冷门主题判断上表现优异

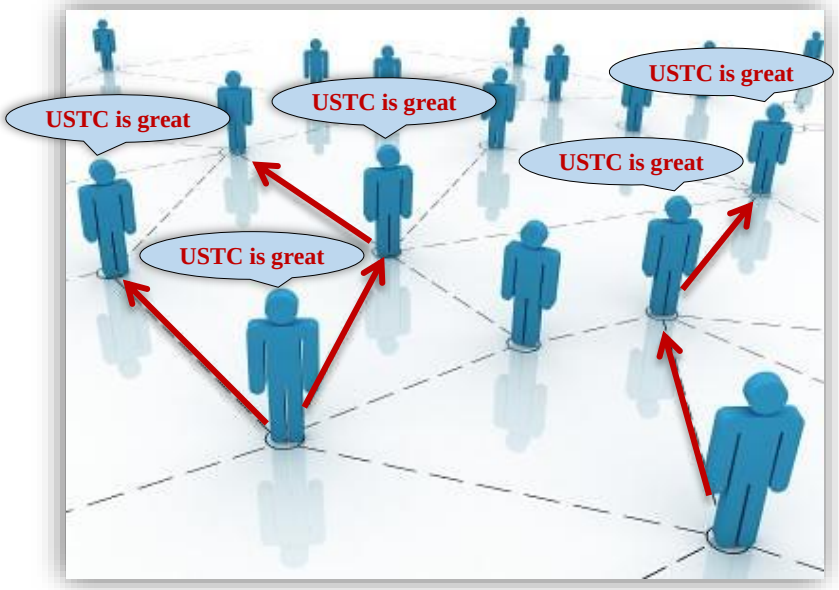
本节目录

- 基本传播模型
  - 独立级联模型
  - 线性阈值模型
- 传播最大化问题
- 衍生传播模型与传播问题
- **基于社会网络的推荐**

本节主旨

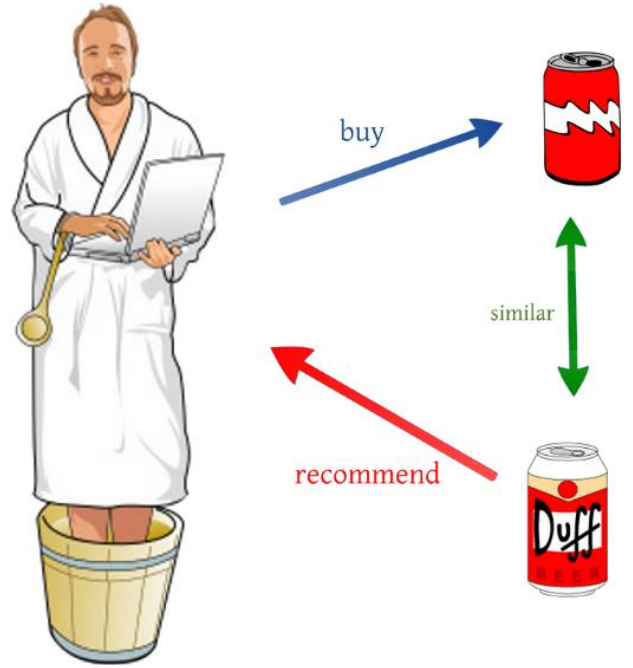
- **从信息级联到社交推荐**

- 信息级联教会我们的道理：人是可以被影响的
- 这种影响将作用于在我们的决策
  - 如何量化描述社交因素对于决策的影响，进而实现有针对性的推荐?
  - 社交因素采用何种方式融入已有的推荐模型?



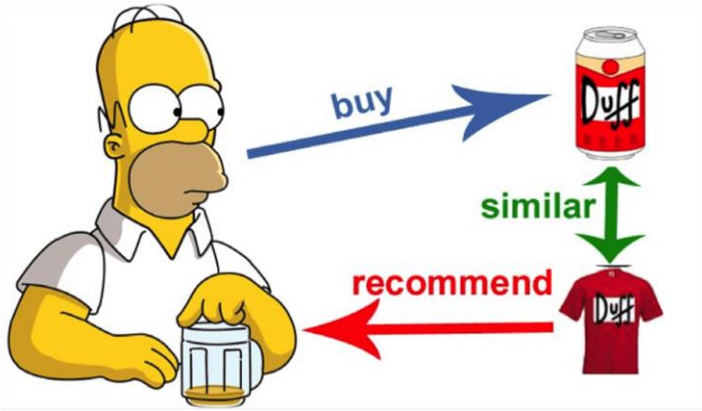
- **推荐问题的核心思想**

- 基本想法：用户的偏好一般相对稳定
  - 因此，给用户推荐他/她以前喜欢的物品准没错（偏保守的想法）
    - 例如：电影推荐
      - 推荐同演员、同导演、同主题……
    - 例如：新闻推荐
      - 推荐类似主题/倾向性的文章
        - 可能造成局限性和错觉



- **基本推荐の优缺点（+/-）**

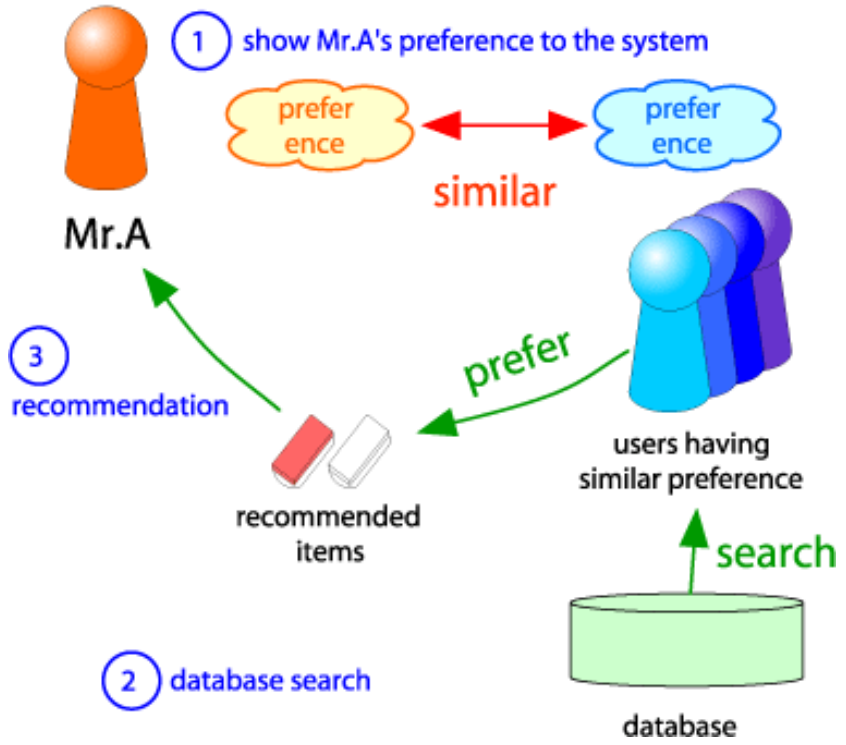
- 每个人的推荐过程相互独立，受冷启动影响较小
- 可以为具有独特偏好的用户进行有效推荐，不受大众倾向性和热度的影响
- 可以推荐新物品或非热门物品
- 找到合适的特征是一件困难的事
  - 对于非结构化信息，如图像、视频、音频等尤其如此
- 给新用户推荐物品，永远是一个困难的任务
- 过度特化（Overspecialization）现象
  - 永远只能给用户推荐局限于画像的内容（小心伦理问题！）
  - 用户的多方面兴趣难以体现



- **协同过滤：破解基本推荐的有趣思路**

- 如何破解基本推荐方案的诸多局限性？

- 在实际应用中，我们发现，其他用户的行为对当前用户有借鉴作用
  - 例如，你和你的狐朋狗友们往往具有相似的口味
  - 相应的，这种相似的口味导致了在选择上的相似性
  - 重要的是，这种借鉴行为不需要对推荐对象进行表征，也能够推荐更多样、更新颖的结果



- **协同过滤的基本思想**

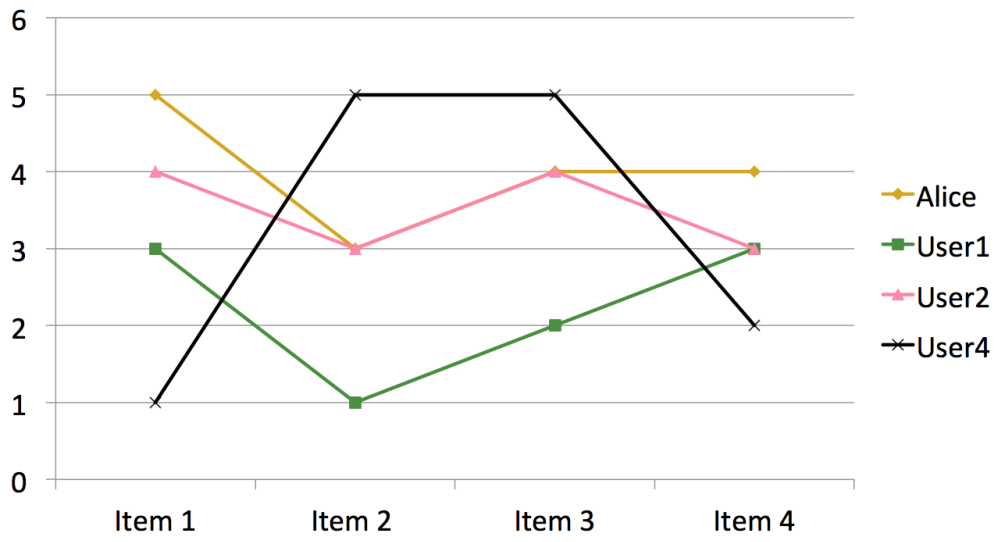
- 如前所述，推荐系统的本质就是矩阵补全问题
- 相应的，协同过滤的思想在于基于矩阵的其他行，协助填补本行的空缺

Target User	Item 1	Item 2	Item 3	Item 4	Item 5	Average
Alice	5	3	4	4	???	16/4
User1	3	1	2	3	3	9/4
User2	4	3	4	3	5	14/4
User3	3	3	1	5	4	12/4
User4	1	5	5	2	1	13/4

- **协同过滤与社交元素的碰撞：基于用户推荐**

- 如前所述，具有相似偏好的用户，往往在过去与未来的评分行为上相似

- 基于用户（User-based）推荐的目的，即在于找到这些相似用户，并基于这些用户的历史行为进行推荐
  - 相似用户往往被称作“邻居”，其本质即没有显式社交关系的“好友”



- **最近邻的寻找过程**

- 如前所述，具有相似偏好的用户，往往在过去与未来的评分行为上相似
- 相应的，寻找最近邻的依据，应该从用户过去的评分行为上着手
  - 历史评分行为越相似，用户之间未来行为的相似性就越高
  - 基于共同评分的物品，衡量用户之间的相似性如下：

$$sim(a,b) = \frac{\sum_{p \in product(P)} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in product(P)} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in product(P)} (r_{b,p} - \bar{r}_b)^2}}$$

Average rating of user ***b***

- **基于用户推荐的评分预测**

- 在得到用户之间的相似性后，针对待预测的物品，可以根据历史上其他用户对于该物品的评分，结合用户之间的相似性作为加权，预测评分结果
  - 相似性计算与评分预测中，都通过减去平均值来抹去个人评分偏好的影响
    - 不同用户打分范围不同，有些人倾向于打高分，有些人更苛刻

$$pred(a,p) = \bar{r}_a + \frac{\sum_{b \in neighbors(n)} sim(a,b) \cdot (r_{b,p} - \bar{r}_b)}{\sum_{b \in neighbors(n)} sim(a,b)}$$

- **基于协同过滤的推荐的优缺点**

- 优点：可适用于任意种类物品，效果较好
  - 不受多模态、非结构化信息表征与特征选取的困扰
- 缺点：稀疏性、热度偏差等
  - 稀疏性：用户评分记录严重稀疏，很难找到评价过同一物品的用户
  - 热度偏差：更倾向于推荐热门物品，对具有独特偏好的用户推荐效果差
    - 寻找相似用户/商品时，小众偏好很容易被热门偏好所淹没
  - 社交可达性：不是你的好友，会对你产生影响吗？

- **第一步：协同过滤的小改进**

- 如前所述，协同过滤的根基是“相似用户”，但相似用户未必有影响的渠道
  - 如何将“社交可达性”与协同过滤连接起来？
  - 可行方法 1：只保留最近邻中具有好友关系的部分

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in N(u) \cap F(u)} sim(u,v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N(u) \cap F(u)} sim(u,v)}.$$

- **第一步：协同过滤的小改进**

- 如前所述，协同过滤的根基是“相似用户”，但相似用户未必有影响的渠道
  - 如何将“社交可达性”与协同过滤连接起来？
    - 可行方法 1 也有隐患，如果最近邻中没有社交好友（不重叠），怎么办？
  - 可行方法 2：按相似性排名，选取前 K 个好友（只在好友中挑选）

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in S(u)} sim(u,v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in S(u)} sim(u,v)}.$$

  - 其中S(u)为按照前述规则挑选的邻居集合（真正的社交邻居）

- **还不够，我们希望更一般化的模型**

- 协同过滤推荐技术，仅对数据进行简单处理，适用于各种数据
- 然而，数据的稀疏性、计算最近邻的高复杂度，限制了其有效性
- 与此同时，我们知道，从效用矩阵的视角来看，推荐系统的本质是矩阵补全
- 那么，矩阵的各个元素是如何生成的？
  - 基本思路：用户对物品的评分，本质上是用户的偏好，与物品的属性之间的相似度。相似度越高，评分越高
  - 那么，用户的偏好与物品的属性，如何表示？

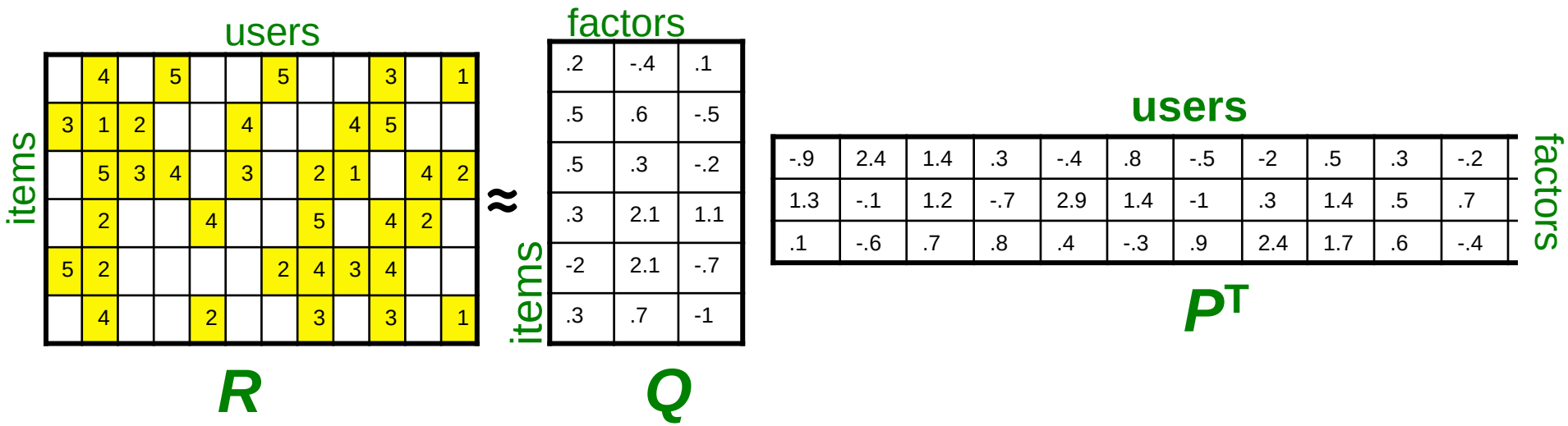
- **还不够，我们希望更一般化的模型**

- 协同过滤推荐技术，仅对数据进行简单处理，适用于各种数据
  - 然而，数据的稀疏性、计算最近邻的高复杂度，限制了其有效性
  - 与此同时，我们知道，从效用矩阵的视角来看，推荐系统的本质是矩阵补全

	Avatar	LOTR	Matrix	Pirates
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

- **基本的矩阵分解推荐思路**

- 借鉴矩阵分解（Matrix Factorization）的思路，揭示潜在因子



	4		5			5			3		1
3	1	2			4			4	5		
	5	3	4		3		2	1		4	2
	2			4			5		4	2	
5	2					2	4	3	4		
	4			2			3		3		1

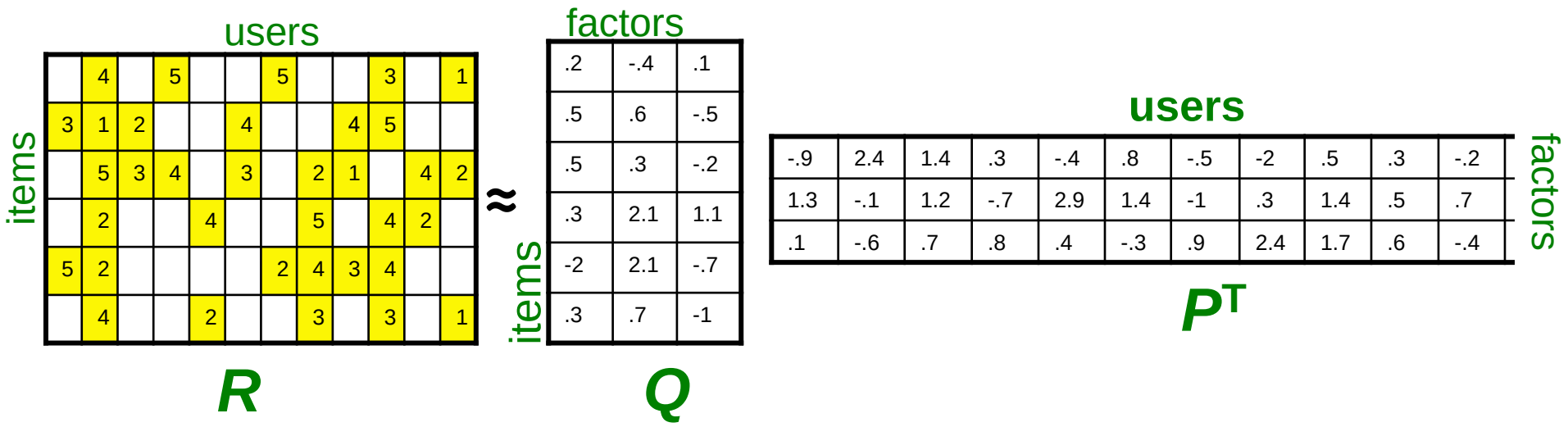
.2	-.4	.1
.5	.6	-.5
.5	.3	-.2
.3	2.1	1.1
-2	2.1	-.7
.3	.7	-1

-.9	2.4	1.4	.3	-.4	.8	-.5	-2	.5	.3	-.2
1.3	-.1	1.2	-.7	2.9	1.4	-1	.3	1.4	.5	.7
.1	-.6	.7	.8	.4	-.3	.9	2.4	1.7	.6	-.4

- 评分矩阵R被近似视作物品属性矩阵Q与用户偏好矩阵P的乘积
- P与Q的维度，一方面与用户/物品的数量有关，另一方面体现了潜在因子的数量
  - 开放问题：Latent Factor的维度如何确定？

- **基于模型的推荐：评分预测**

- 借鉴矩阵分解（Matrix Factorization）的思路，揭示潜在因子



	4		5			5			3		1
3	1	2			4			4	5		
	5	3	4		3		2	1		4	2
	2			4			5		4	2	
5	2					2	4	3	4		
	4			2			3		3		1

.2	-.4	.1
.5	.6	-.5
.5	.3	-.2
.3	2.1	1.1
-2	2.1	-.7
.3	.7	-1

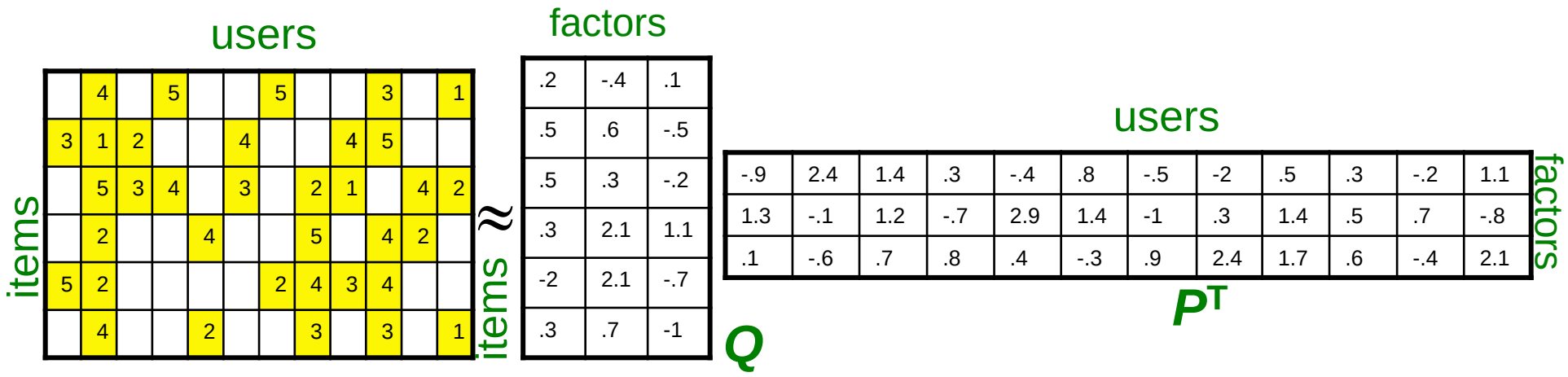
-.9	2.4	1.4	.3	-.4	.8	-.5	-2	.5	.3	-.2
1.3	-.1	1.2	-.7	2.9	1.4	-1	.3	1.4	.5	.7
.1	-.6	.7	.8	.4	-.3	.9	2.4	1.7	.6	-.4

- 当用户与物品的潜在因子已知，则任何缺失的评分，均可以通过对应的P、Q矩阵相应的行列运算估计得到

$$\hat{r}_{xi} = q_i \cdot p_x = \sum_k q_{ik} \cdot p_{xk}$$

- **基于模型的推荐：潜在因子**

- 那么，如何估计出两个潜在因子矩阵?
  - 既然用户评分是根据潜在因子的乘积所得，那么，基于这种方法得到的用户评分，应与历史评分记录尽可能接近，即 $\min_{P,Q} \sum_{(i,x)\in R} (r_{xi} - q_i \cdot p_x)^2$



- **千呼万唤始出来：将社交元素引入矩阵分解**

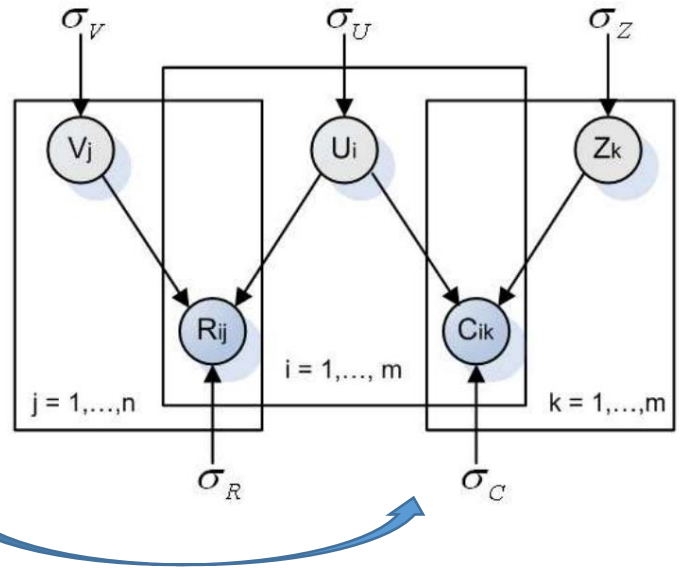
- 前面我们介绍了基于矩阵分解的推荐技术，然而，如何引入社交元素？
  - 回到开头的基本假设：狐朋狗友之间有着相似的口味
  - 体现在“潜在因子”上，就是好友之间有着相似的向量表示
  - 因此，通过在潜在因子上加上社交约束的方式，可以实现基于社交关系的推荐
  - 一句题外话：曾经盛行的三大挖坑方式
    - 行内的黑话：给概率图加圈，给矩阵分解加约束，给神经网络加层

• 矩阵分解的拓展：社交约束

- 最简单的社交约束是什么样子？
  - 直观的想法：社交影响仅作用于用户属性，通过影响属性间接影响行为
  - 因此，如果我们知道好友关系作为Side Information，可以对模型进行补充
  - 一个基于社交约束方面的经典工作：SoRec
    - 右半边表示社交约束，C表示社交关系
      - 存在社交关系的，属性必然相似

$$p(C|U,V,\sigma_R^2)=\prod_{i=1}^{m}\prod_{j=1}^{n}\mathcal{N}\left[\left(r_{ij}|g(U_i^TV_j),\sigma_R^2\right)\right]^{I_{ij}^R}$$

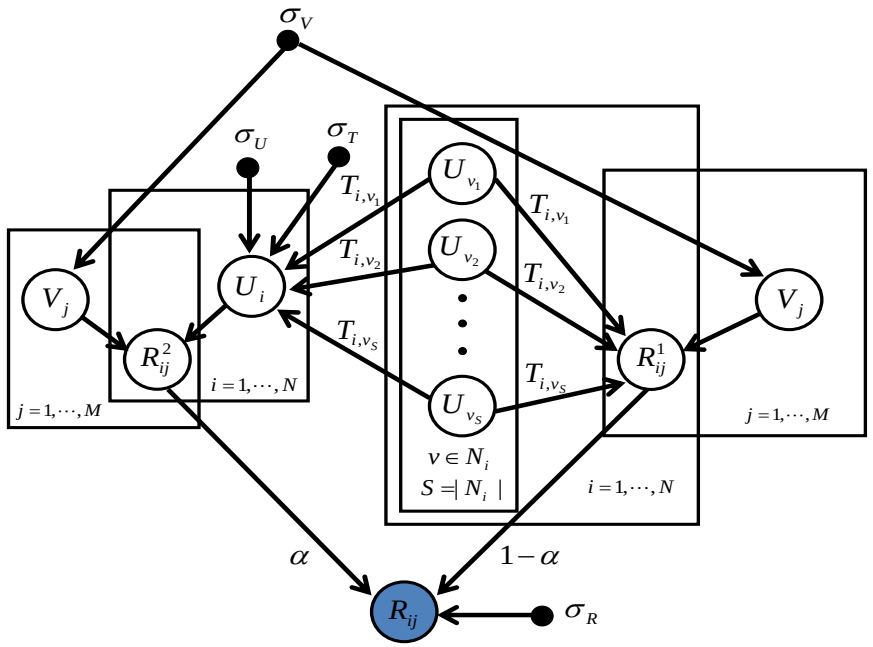
$$p(C|U,Z,\sigma_C^2)=\prod_{i=1}^{m}\prod_{k=1}^{m}\mathcal{N}\left[\left(c_{ik}|g(U_i^TZ_k),\sigma_C^2\right)\right]^{I_{ik}^C}$$



H Ma, et al., SoRec: social recommendation using probabilistic matrix factorization, CIKM 2008

• 矩阵分解的拓展：社交约束

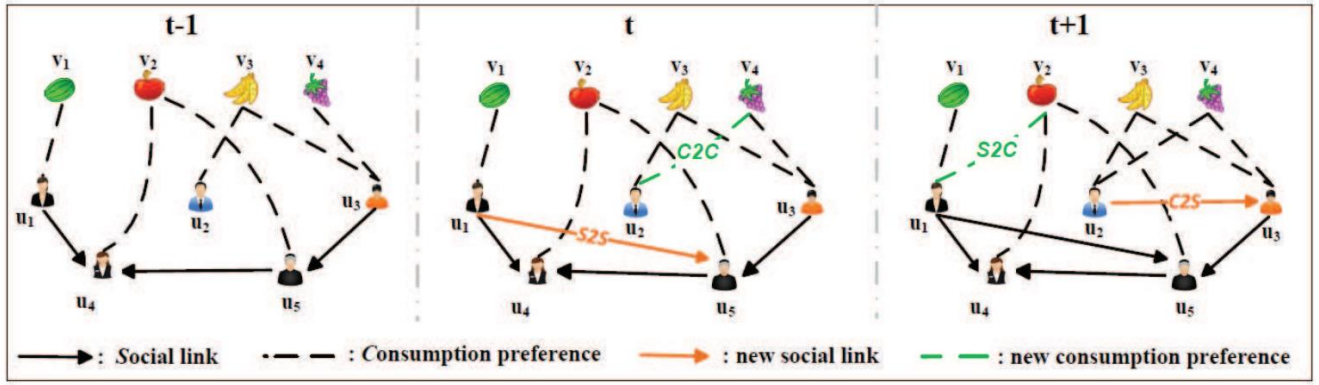
- 基于属性相似的社交约束虽然效果很好，但未必始终成立
  - 例如：社交关系可能并非通过偏好间接影响决策，而是直接影响决策
  - 右图的框架同时体现了这两种思路
    - R1为好友直接影响评分
    - R2为好友通过影响属性间接影响评分



杜东舫等，基于信任机制下概率矩阵分解的用户评分预测，软件学报，2018, 29(12):3747-3763.

- **更进一步：社交与行为的联合演化**

- 先前所介绍的所有社交推荐问题，其本质都是建立在固定的社交网络上
  - 然而现实中，社交网络的结构在不断发生演化（社交选择与社交影响交互作用）
  - 在上一节课中我们讨论过：社交影响与社交选择互为因果，从而不断强化节点同质性
    - 是否可以将这一现象体现在社交推荐建模中？



Explanations of A2B

A2B	Explanations
C2C	Collaborative Filtering
S2C	Social Influence
S2S	Node Proximity
C2S	Homophily Effect

- **更进一步：社交与行为的联合演化**

- 将关系-行为相互作用转化为时间轴上的交叉演进，实现行为预测与社交推荐

演化消费行为

$$p(U_a^t) = \mathcal{N}(U_a^t | \bar{U}_a^t, \sigma_U^2 \mathbf{I})$$

$$\text{where} \quad \bar{U}_a^t = (1-\alpha_a)U_a^{(t-1)} + \alpha_a \sum_{b \in N_a^{(t-1)}} F_{ab}^t U_b^{(t-1)}$$

$$s.t. \quad \forall a \in U, \quad 0 \le \alpha_a \le 1, \qquad (3)$$

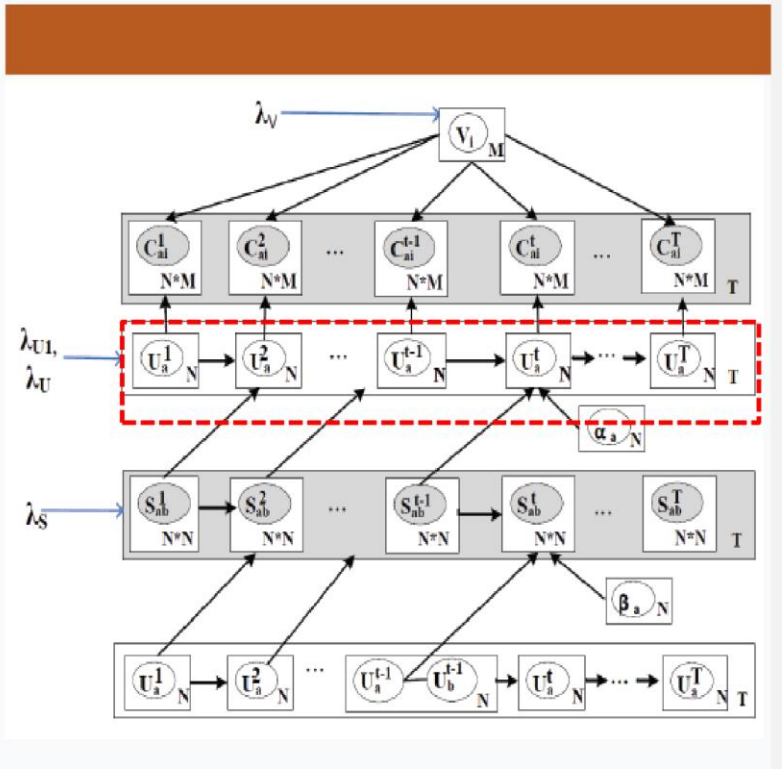
社交影响力

演化社交行为

$$\hat{S}_{ab}^t = (1-\beta_a)h(a,b,t) + \beta_a \{U_a^{(t-1)}, U_b^{(t-1)}\},$$

同质性

$$h(a,b,t) = \frac{1}{\sum_{c \in N_a^{(t-1)} \cap N_b^{(t-1)}} log(|N_c^{(t-1)}|)}.$$



Le Wu, et. al., Modeling Users' Preferences and Social Links in Social Networking Services: a Joint-Evolving Perspective, AAAI 2016

本章小结

经典传播模型

- 基本传播模型
  - 独立级联模型
  - 线性阈值模型
- 传播最大化问题
- 衍生传播模型与传播问题
- 基于社会网络的推荐

tongxu@ustc.edu.cn